BOOMBIKES ASSIGNMENT

Model to predict demand

Abstract

A US based bike sharing model – BoomBikes wants to understand the factors on which the demand for shared bikes depends. This assignment is carried out to build a model for such prediction. The document covers theorical aspects of linear regression

Amarjit Mahadik

Amarjit.mahadik@gmail.com

Name: Amarjit Mahadik

Assignment Name: BoomBikes Linear Regression Model to predict the demand for bike rides.

Submission Date: 07-Jan-24

Assignment-based Subjective Questions:

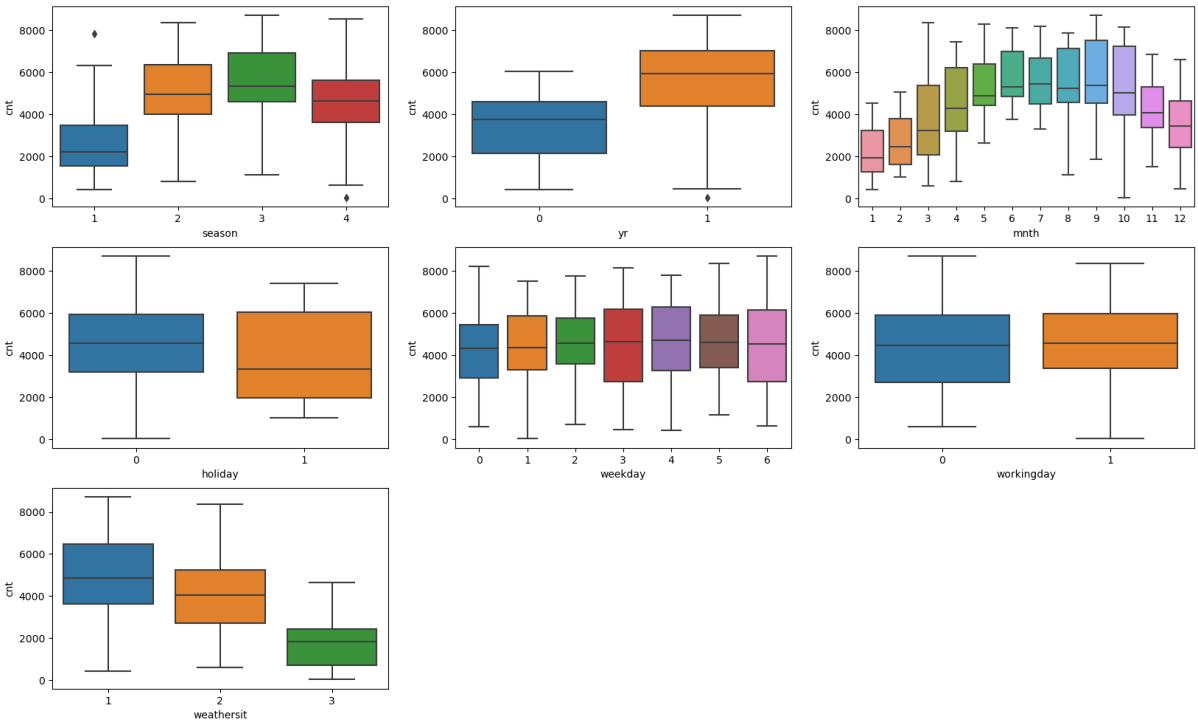
Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

There are 7 categorical variables - 'season','yr','mnth','holiday','weekday','workingday','weathersit'. I analyzed the impact on dependent variable using box plot. The inferences that I got are as below:

- Demand for bikes seems higher in Season 2 (Summer)and 3 (fall)
- Demand in 2019 is higher
- Season and Month are corelated
- Demand is higher when the weather is clear

The box plot is as below:



Question 2: Why is it important to use drop_first=True during dummy variable creation?

Answer:

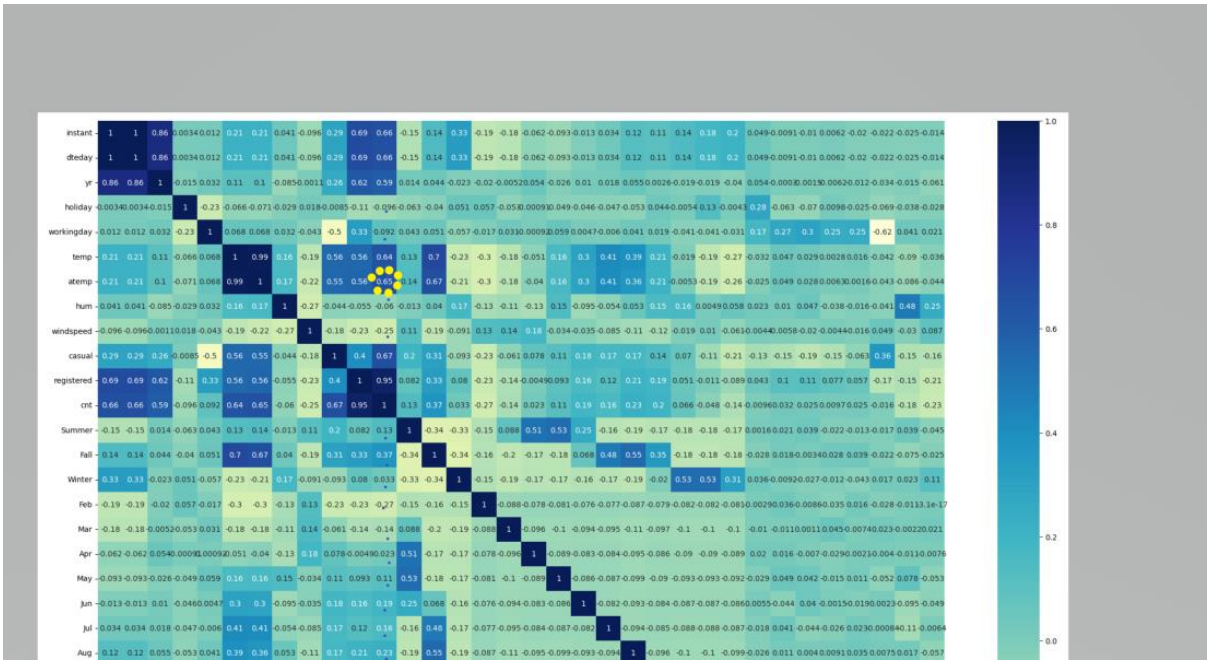
The common approach to handling categorical variables is to convert each category value into a new column and assigning a 1 or 0 (True/False) value to the column. This is particularly useful for nominal variables where there is no inherent order. For example, season can be split into four columns - one for each season - with a 1 or 0 indicating the season of the year. However this can lead to multi collinearity due to addition of multiple columns. To avoid this one level of categorical variable can be dropped as t is redundant.

In dummy variable creation, drop_first = True ensures that the first dummy variable is dropped.

Question 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:

The feature "atemp - feeling temperature in Celsius" has the highest corelation (0.65) with target variable.

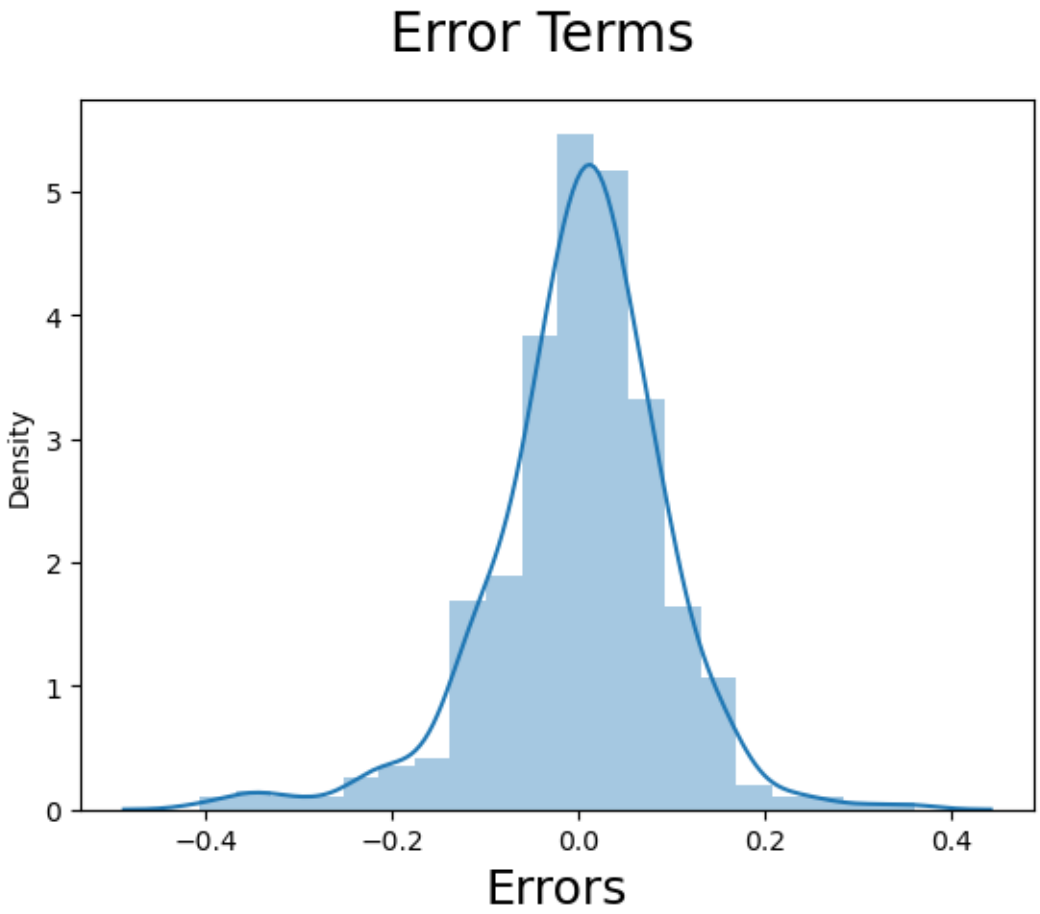


Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

Below mentioned are the assumptions of linear regression. I have also mentioned validation technique:

- Normality : It is assumed that the error terms ,are normally distributed.
- Multicollinearity : Variable must be independent of each other . To validate this, we checked the VIF of the features. The features with high value (>5) were drpped.
- Homoscedasticity : It is assumed that the residual terms have the same (but unknown) variance. So, now to check if the error terms are also normally distributed, we lotted the histogram of the error terms and see what it looks like.
- Autocorrelation : Autocorrelation is the correlation between two of the same series. This is mainly used in time series data. Since ours is not time series data, we ignored it.



Question 5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:
As we can see the model for predicting the demand is as below

$$Rented_Bikes = 0.493 \times atemp + 0.232 \times yr - 0.138 \times hum - 0.164 \times windspeed - 0.055 \times okweather - 0.237 \times badweather + 0.124 \times Summer + 0.0967 \times Fall + 0.164 \times Winter - 0.0834 \times holiday$$

The top 3 variables for predicting the demand are:

1. *Feeling Temperature:* Temperature is positively corelated with demand. The higher the temperature, higher the demand.
2. *Bad weather:* This variable is neh=gatively corelated with demand. If there is rain or snow the demand is lower.
3. *Year:* This variable is positively corelated and hence we see the demand is higher in 2019. We can safely assume that the trend for shared bikes is picking up and is supposed to go higher as we move out of pandemic.

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer:

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. The basic idea is to fit a linear equation to observed data. Here's a detailed explanation:

1. *Model Representation:* In linear regression, the relationship between the dependent variable Y and independent variables X is modeled as Y = Intercept + Coef_1 * X_1 + Coef_2 * X_2 + ... + Coef_n * X_n + Error_term
2. *Best Fit Line:* The algorithm tries to find the best fit line which minimizes the difference between the predicted values and actual values. This difference is known as the residual.
3. *Least Squares Method:* It commonly uses the least squares method to minimize the sum of the squared residuals, providing the 'line of best fit'.
4. *Coefficients Estimation:* The coefficients are estimated during the training process, using statistical methods like Ordinary Least Squares (OLS).
5. *Prediction:* Once the model is trained and coefficients are estimated, it can be used to predict the dependent variable for given values of independent variables.
6. *Assumptions:* Linear regression assumes linearity, independence, homoscedasticity, and normal distribution of residuals.
7. *Single vs Multiple Regression:* In simple linear regression, there is one independent variable, while in multiple linear regression, there are multiple independent variables.
8. *Interpretation:* The coefficients represent the change in the dependent variable for a one-unit change in an independent variable, assuming all other variables are held constant.
9. *Diagnostics and Validation:* After model fitting, it's important to validate assumptions using diagnostic plots and tests for residuals.
10. *Application:* It's widely used in various fields for predictive modeling, trend analysis, and causal inference, given its simplicity and interpretability.

Question 2: Explain the Anscombe's quartet in detail

Answer:

Anscombe's quartet comprises four different datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven points and was constructed in 1973 by the statistician Francis Anscombe to demonstrate the importance of graphical representation of data when analyzing statistical relationships. Despite having the same mean, variance, correlation, and linear regression line ($y = 3.00 + 0.500x$) when applied to each dataset, the distributions are fundamentally different, as revealed through their scatter plots. This quartet illustrates the pitfalls of relying solely on summary statistics and highlights the need for visualizing data to uncover underlying data structures. Anscombe's work emphasizes that statistical analysis should go beyond numerical calculations to include graphical analysis, which can reveal underlying patterns or anomalies that numbers alone might miss.

Question 3: What is Pearson’s R?

Answer:

Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that describes the strength and direction of a linear relationship between two continuous variables. It is a value between -1 and 1 where:

- +1 indicates a perfect positive linear correlation:** As one variable increases, the other variable increases at a constant rate.
- -1 indicates a perfect negative linear correlation:** As one variable increases, the other decreases at a constant rate.
- 0 indicates no linear correlation:** There is no linear relationship between the two variables.

Pearson's correlation coefficient is widely used in the sciences and social sciences to assess the degree of linear relationship between two variables, often as a preliminary step to more detailed analysis. It's important to note that while Pearson's R can inform about the strength and direction of a linear relationship, it does not indicate causation.

Question 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

Scaling is a data preprocessing technique used to standardize the range of independent variables or features of data. It's important in scenarios where the data comprises features with varying scales, as these differing scales can distort the relative contributions of each feature when modeling.

Why Scaling is Performed:

- *To Avoid Bias:* In many machine learning algorithms, especially those involving distance calculations (like k-NN and SVM) or gradient descent optimization (like linear regression and neural networks), features with larger scales can disproportionately influence the model.
- *To Improve Convergence:* Algorithms converge faster when features are on a similar scale.
- *To Enhance Interpretability:* It helps in comparing coefficients in models where the features are of different units.

Normalized Scaling vs Standardized Scaling:

1. *Normalized Scaling (Min-Max Scaling):* This rescales the feature to a fixed range, usually 0 to 1. The formula is (X - X_min) / (X_max - X_min)). It's useful when you need to bound values but can be sensitive to outliers.
2. *Standardized Scaling (Z-score Normalization):* This rescales the feature so that it has a mean of 0 and a standard deviation of 1. The formula is ((X – X_bar) / sigma), where X_bar is the mean and sigma is the standard deviation. This method is less affected by outliers and is often used when the algorithm assumes the input data to be normally distributed.

In summary, scaling is essential for harmonizing the scales of different features in a dataset, with normalization and standardization being two common techniques, each useful in different contexts and for different types of data.

Question 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

The Variance Inflation Factor (VIF) is a measure used to detect the presence and severity of multicollinearity in a regression analysis. It quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated. A VIF value can sometimes be infinite, and this typically occurs due to one of the following reasons:

- *Perfect Multicollinearity:* This happens when one independent variable is an exact linear combination of another. For instance, if you have two variables where one is always twice the value of the other, they are perfectly collinear. In such cases, the denominator in the VIF formula becomes zero, leading to an infinite VIF.
- *Highly Correlated Variables:* While perfect multicollinearity is a clear-cut case, VIF can also approach infinity with high but not perfect multicollinearity. If two variables are almost perfectly correlated (e.g., a correlation coefficient very close to 1 or -1), the VIF for these variables will be extremely high, potentially approaching infinity as the correlation approaches 1.
- *Redundant Variables:* Including variables in the model that are unnecessary or redundant can lead to high multicollinearity. For example, including dummy variables for all categories of a categorical variable without omitting one category (the reference category) can cause one of the dummy variables to be perfectly predicted by the others, resulting in an infinite VIF.
- *Data Sampling Issues:* Sometimes, the way data is sampled or collected can result in an artificial multicollinearity, which can lead to high or infinite VIF values.

An infinite VIF is a clear indication that your regression model has multicollinearity issues that need to be addressed, either by removing or combining variables, or by using regularization techniques that can handle multicollinearity.

Question 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to compare two probability distributions by plotting their quantiles against each other. In the context of linear regression, it is typically used to assess whether a dataset follows a certain theoretical distribution, usually a normal distribution.

Use and Importance of a Q-Q Plot in Linear Regression:

- *Assessing Normality of Residuals:* In linear regression, one of the key assumptions is that the residuals (the differences between the observed values and the values predicted by the model) are normally distributed. A Q-Q plot is used to visually check this assumption. The residuals of the model are plotted against a normal distribution in such a way that if the residuals are normally distributed, the points should fall approximately along a straight line.
- *Identifying Skewness and Kurtosis:* A Q-Q plot can also help in identifying deviations from normality like skewness (where the distribution is stretched to one side) and kurtosis (the "peakedness" of the distribution). If the points in a Q-Q plot deviate from the straight line in a systematic way, it indicates that the residuals have a distribution that is skewed or has high kurtosis.
- *Diagnosing Model Fit:* The plot can be used to diagnose whether a linear model is a good fit for the data. Significant deviations from the straight line in the Q-Q plot suggest that the model may not be capturing some aspects of the data's structure, which might mean that a non-linear model is more appropriate.
- *Improving Model Accuracy:* By identifying deviations from normality in the residuals, a Q-Q plot can guide the transformation of variables. For instance, if residuals are right-skewed, applying a log transformation might improve the model.

In summary, Q-Q plots are crucial for verifying one of the key assumptions of linear regression (normality of residuals), and they play an important role in model diagnostics and validation. A well-fitting linear regression model should have residuals that approximately follow a straight line in the Q-Q plot. Deviations from this line indicate potential problems with the model's assumptions or suggest the need for data transformation.